Event

OrionX 2017 Data Center Predictions

A variation of this article was originally published in The Register, "Wow, what an incredible 12 months: 2017's data center year in review", Feb 6, 2017.

The data center market is hot, especially now that we are getting a raft of new stuff, from promising non-Intel chips and system architectures to power and cooling optimizations to new applications in Analytics, IoT, and Artificial Intelligence.

Here is our Top-10 Data Center Predictions for 2017:

1) Data center optimization is here

Data centers are information factories with lots of components and moving parts. There was a time when companies started becoming much more complex, which fueled the massive enterprise resource planning market. Managing everything in the data center is in a similar place. To automate, monitor, troubleshoot, plan, optimize, cost-contain, report, etc, is a giant task, and it is good to see new apps in this area.

Data center infrastructure management provides visibility into and helps control IT. Once it is deployed, you'll wonder how you did without it. Someday, it will be one cohesive thing, but for now, because it's such a big task, there will be several companies addressing different parts of it.

2) Azure will grow faster than AWS

Cloud is the big wave, of course, and almost anything that touches it is on the right side of history. So, private and hybrid clouds will grow nicely and they will even temper the growth of public clouds. But the growth of public clouds will continue to impress, despite increasing recognition that they are not the cheapest option, especially for the mid-size users.

AWS will lead again, capturing most new apps. However, Azure will grow faster, on the strength of both landing new apps and also bringing along existing apps where Microsoft maintains a significant footprint.

Moving Exchange, Office, and other apps to the cloud, in addition to operating lots of regional data centers and having lots of local feet on the ground will help.

Some of the same dynamics will help Oracle show a strong hand and get close to Google and IBM. And large telcos will stay very much in the game.

Smaller players will persevere and even grow, but they will also start to realize that public clouds are their supplier or partner, not their competition! It will be cheaper for them to OEM services from bigger players, or offer joint services, than to build and maintain their own public cloud.

3) Great Wall of persistent memory will become a thing

Just as the hottest thing in enterprise becomes Big Data, we find that the most expensive part of computing is moving all that data around. Naturally, we start seeing what OrionX calls "In-Situ Processing" (see page 4 of the report in the link): instead of data going to compute, compute would go to data, processing it locally wherever data happens to be.

But as the gap between CPU speed and storage speed separates apps and data, memory becomes the bottleneck. In comes storage class memory (mostly flash, with a nod to other promising technologies), getting larger, faster and cheaper. So, we will see examples of apps using a Great Wall of persistent memory, built by hardware and software solutions that bridge the size/speed/cost gap between traditional storage and DRAM. Eventually, we expect programming languages to naturally support byte-addressable persistent memory.

4) System vendors will announce racks, not servers

Vendors already configure and sell racks, but they often populate them with servers that are designed as if they'd be used stand-alone. Vendors with rack-level thinking have been doing better because designing the rack vs the single node lets them add value to the rack while removing unneeded value from server nodes.

So, server vendors will start thinking of a rack, not a single-node, as the system they design and sell. Intel's Rack Scale Architecture has been on the right track, a real competitive advantage, and an indication of how traditional server vendors must adapt. The server rack will become the next level of integration and what a "typical system" will look like. Going forward, multi-rack systems are where server vendors have a shot at adding real value. HPC vendors have long been there.

5) Server revenue growth will be lower than GDP growth

Traditional enterprise apps – the bulk of what runs on servers – will show that they have access to enough compute capacity already. Most of that work is transactional, so their growth is correlated with the growth in GDP, minus efficiencies in processing.

New apps, on the other hand, are hungry, but they are much more distributed, more focused on mobile clients, and more amenable to what OrionX calls "High-Density Processing": algorithms that have a high ops/bytes ratio running on hardware that provides similarly high ops/byte capability – ie, compute accelerators like GPUs, FPGAs, vector processors, manycore CPUs, ASICs, and new chips on the horizon.

On top of that, there will be more In-Situ Processing: processing the data wherever it happens to be, say locally on the client vs sending it around to the backend. This will be made easier by

the significant rise in client-side computing power and more capable data center switches and storage nodes that can do a lot of local processing.

We will also continue to see cloud computing and virtualization eliminate idle servers and increase the utilization rates of existing systems.

Finally, commoditization of servers and racks, driven by fewer-but-larger buyers and standardization efforts like the Open Compute Project, put pressure on server costs and limit the areas in which server vendors can add value. The old adage in servers: "I know how to build it so it costs $1m, but don't know how to build it so it's worth $1m" will be true more than ever.

These will all combine to keep server revenues in check. We will see 5G's wow-speeds but modest roll-out, and though it can drive a jump in video and some server-heavy apps, we'll have to wait a bit longer.

6) OpenPOWER will emerge as the viable alternative to x86

The real battle in server architecture will be between Intel's in-house coalition and what my colleague Dan Olds has called the Rebel Alliance: IBM's OpenPower industry coalition. Intel brings its all-star team: Xeon Phi, Altera, Omni-Path (plus Nervana/Movidius) and Lustre, while OpenPower counters with a dream team if its own: POWER, Nvidia, Xilinx, and Mellanox (plus TrueNorth) and GPFS (Spectrum Scale). Then there is Watson which will become more visible as a differentiator, and a series of acquisitions by both camps as they fill various gaps.

The all-in-house model promises seamless integration and consistent design, while the extended team offers a best-of-breed approach. Both have merits. Both camps are pretty formidable. And there is real differentiation in strategy, design, and implementation. Competition is good.

7) Beware IoT Security

One day in the not too distant future, your fancy car will break down on the highway for no apparent reason. It will turn out that the auto entertainment system launched a denial of service attack on the rest of the car, in a bold attempt to gain control. It even convinced the nav system to throw in with it! This is a joke, of course, but could become all too real if/when human and AI hackers get involved.

IoT is coming, and with it will come all sorts of security issues. Do you know where your connected devices are? Can you really trust them?

8) More chips than Vegas; riskier too

For the first time in decades, there is a real market opening for new chips. What drove this includes:

- The emergence of new apps, led by AI and IoT. The part of AI that is computationally interesting and different is High-Performance AI, since it intersects with HPC. On the

IoT side, backend apps are typically Analytics to make sense of sensor data. These new apps will run where they can run better/faster/cheaper. They will be too new to be burdened by any allegiance or bonds to a particular chip.

- The fact that many existing apps have no clue what hardware they run on, and operate on the upper layers of a tall stack.
- The possibility to build a complete software stack from open source software components.
  The presence of very large customers like cloud providers or top supercomputing sites. They buy in seriously large volumes and have the wherewithal to build the necessary software stack, so they can afford to pick a new chip and bolster, if not guarantee, its viability.

This will be a year when many new chips became available and tested, and there is a pretty long list of them, showing just how big the opportunity is, how eager investors must have been to not miss out, and how many different angles there are.

In addition to AI, there are a few important general-purpose chips being built. The coolest one is by startup Rex Computing, which is working on a chip for exascale, focused on very high compute-power/electric-power ratios. Qualcomm and Caviuum have manycore ARM processors, Oracle is advancing SPARC quite well, IBM's POWER continues to look very interesting, and Intel and AMD push the X86 envelope nicely.

With AI chips, Intel already has acquired Nervana and Movidius. Google has its TPU, and IBM its neuromorphic chip, TrueNorth. Other AI chip efforts include Mobileye (the company is being bought by Intel since this article was written), Graphcore, BrainChip, TeraDeep, KnuEdge, Wave Computing, and Horizon Robotics. In addition, there are several well-publicized and respected projects like NeuRAM3, P-Neuro, SpiNNaker, Eyeriss, and krtkl going after different parts of the market. That’s a lot of chips, but most of these bets, of course, won’t pay off.

9) ARM server market share will stay below 3 per cent

Speaking of chips, ARM servers will remain important but elusive. They will continue to make a lot of noise and point to significant wins and systems, but fail to move the needle when it comes to revenue market share in 2017.

As a long-term play, ARM is an important phenomenon in the server market – more so now with the backing of SoftBank, a much larger company apparently intent on investing and building, and various supercomputing and cloud projects that are great proving grounds.

But at the end, you need differentiation and ARM has the same problem against X86 as Intel’s Atom had against ARM. Atom did not differentiate vs ARM any more than ARM is differentiating vs Xeon.

Most systems end up being really good at something, however, and there are new apps and an open-source stack to support existing apps, which will help find specific workloads where ARM implementations could shine. That will help the differentiation become more than "it's not X86".

10) Is it an app, or is it a fabric? More cloud fabrics introduced

What is going on with big new apps? They keep getting more modular (microservices), more elastic (scale out), and more real-time (streaming). They've become their own large graph, in the computer science sense, and even more so with IoT (sensors everywhere plus In-Situ Processing).

As apps became services, they began resembling a network of interacting modules, depending on each other and evolving together. When the number of interacting pieces keeps increasing, you've got yourself a fabric. But as an app, it's the kind of fabric that evolves and has an overall purpose (semantic web).

Among engineering disciplines, software engineering already doesn't have a great reputation for predictability and maintainability. More modularity is not going to help with that.

But efforts to manage interdependence and application evolution have already created standards for structured modularity like the OSGi Alliance for Java. Smart organizations have had ways to reduce future problems (technical debt) from the get-go. So, it will be nice to see that type of methodology get better recognized and better generalized.